Keywords- Based on Arabic Information Retrieval Using Light Stemmer

Mohammad Khaled A. Al-Maghasbeh[-1], Mohd Pouzi Bin Hamzah[-2],
[1, 2]-School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu, Kuala Terengganu, Malaysia

Abstract

Arabic morphology complex is a core challenge of Arabic information retrieval. This limitation makes Arabic language is so complex environment of information retrieval specialists due to the high inflected in Arabic language. This paper explains the importance of the stemming in information retrieval through developed a method to search about information needs in Arabic text. The new developed method based on light stemmer to increase matching the keywords between the query, with the related document in the test collection.

Keywords
Arabic Information retrieval, keyword-searching, light stemmer, stemming, Arabic-morphology, natural languages processing.

Introduction

There are many researches have been conducted in information retrieval, question answering, Arabic morphology, and light stemming due to Arabic language represents inflectional and derivational language [7]. Identify the original of word needs to remove prefix, suffix, and other connected character. This process is known as a stemming task. Arabic language has a complexity morphology that makes NLP applications in Arabic language such as information retrieval is very difficult [8]

In addition, stemming is a morphology process to reduce the derived forms of the word. It helps to find the basic or origin of the specific word (root). The root of specific word is the origin of word. Root, also is a basic characters of word without suffix, and prefix. Light Stemming is a process to derivate the roots of each word that written in natural language (NL) such as Arabic text. In other words, it represents that task to generate the morphological form of the word by removing the all diacritics, suffix, and prefix [3][4].

This paper is organized as follows. Section 2 briefly describes the related works. Section 3 describes information retrieval. Section 4 about proposed system. Section 5 discuss and experimental results. In section 6 it summarizes of the work and future work.

2. Related work

The Arabic language is one of the complex languages in the world. Therefore, it needs to special tools and models to deal with its morphology. In study that conducted by Elabd, E., et al, a new approach is applied to deal with information retrieval in Arabic language through analyze all challenges that face the most current used method in information retrieval systems such as latent semantic indexing (LSI), Latent analysis indexing (LAI), Boolean model, and others for attempting to solve them. In the developed approach, also, the query was processed via divide it into two type; the first one is the query that contains only one word, and second one it contains multiple words. In the first case, the query has been stemmed in preparation to match with all documents which consists this word.

However, on the other side of type of query, the stop-words was removed from the phrase, after that stemming all words to match them with related synonyms with all documents that contains at least one of these words [5]. Samy. et al in their paper indicated medical words extraction in multilingual medical resources as a case study of information extraction. The extraction operation has been done by taking the newswire in health field as a dataset sample to attempt compared them with medical list of common Arabic word in Latin prefix and suffix, where as a several tests

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

were conducted by applying several samples, after that, the results were good [10]. Al-Taani. et al presented a new method to parse the Arabic simple sentences using a context free grammar (CFGs) for attempting to remove the ambiguity of Arabic language grammars and to enrich researches of' the natural language processing (NLP) field with a computation system. This method carried out to test 70 of different simple Arabic statements through converts the statements and words into production rules, whereas the results were very good for all tested statements [4].

Abdelali. et al built a project by using cross-languages information retrieval (CLIR) approach that represent a bilingual method. That method aimed to match between the language of the query and the language of the target to facilitate the desired retrieval [1]. Haav. et al, proposed method known as keyword- based information retrieval, whereas these methods currently are extensively used in web search engine. As a result, this type of information retrieval method has a lack for retrieving and fetching all relative information [6].

The common Arabic search engine depends on the keyword in searching about the answer of user's query. Using the web semantic improved the search operation through adds a new layer into the current web that contains a Meta data (or data about data) to related some concepts with each other. Al-khalifa. et al., focused in their search of analyzing the semantic relation among concepts using intelligent characteristics; but there are some challenges in web search understandability of all complex concept, and deep knowledge inside the Arabic textual documents [2]. Study of Noordin. et al., presented a project that designed to retrieve information and versus from Holy Quran to facilitate discover and acquits the knowledge from the Quranic texts [9].

In study of Vallet. et al, a new method was proposed to search about some document in web by using an ontology – based knowledge base to increase and improve the accuracy of search. It's being done through used the related concepts, and synonyms to represent the knowledge, and the build the knowledge base to facilitate of retrieval the correspond documents of queries [11].

3. Information retrieval

The information retrieval includes a lot of methods that used of retrieval. Some of these methods depend on the keywords in search about document or any information through match the query-keywords with the relevant information [6]. So improves the performance of information retrieval systems represents an important task for the majority of the information retrieval researchers [11]. It has recently spread too many studies about the passage retrieval as a branch of information retrieval fields. The passage retrieval is one of information retrieval tasks, which it refers to retrieve a portion of the document. There are some passage retrieval methods such as support victor machine, mixture of language model, and other, which majority of them deals with the frequency or density of required word in each passage to compute the relative ranking of relevant passages [12].

4. Proposed system

The proposed system contains several phase. It starts from preprocessing phase that include normalization tokenization, and stop-word removal. After that using light stemming the terms in both documents, and query. The next step index the document keywords, finally matching between the query-keywords with indexed-terms. These phase will be in briefly explain as follow.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

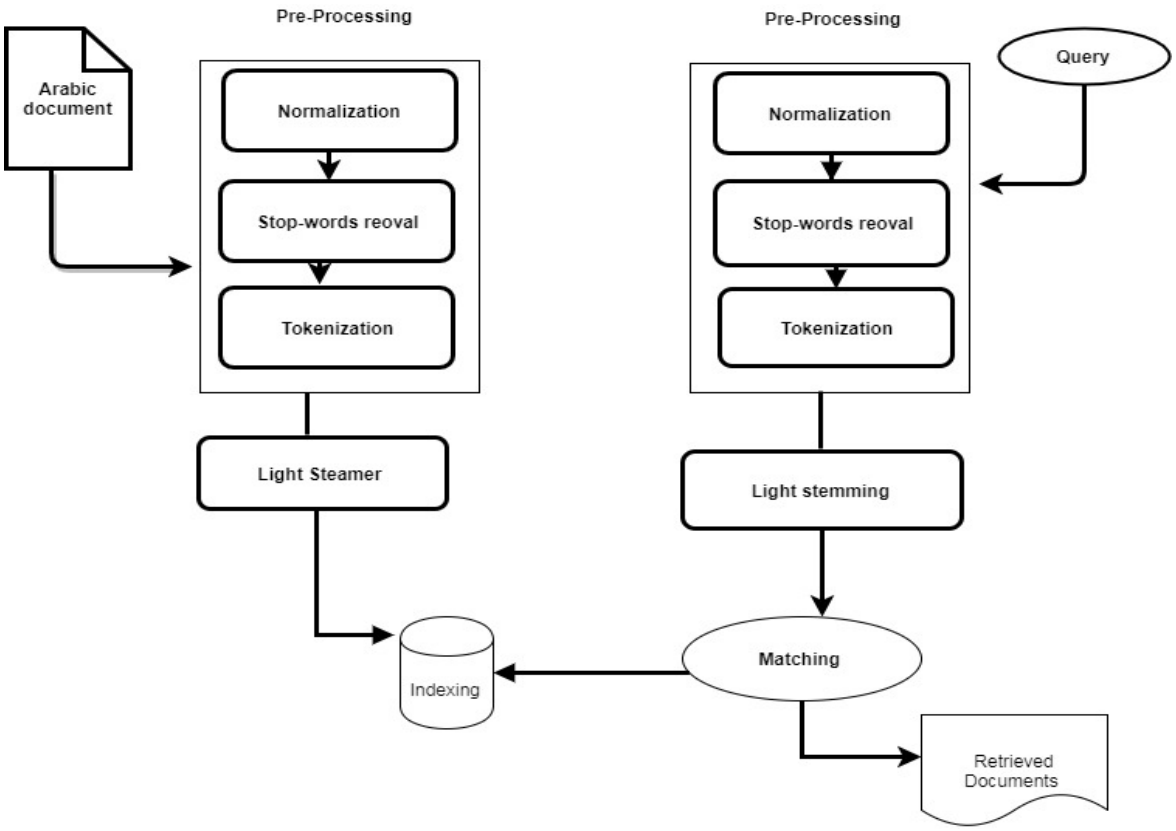

Figure 1: Proposed system architecture

4.1 System phases

4.1.1 Preprocessing phase

This phase is an important process in both, document, and query. The input of this phase is text of modern standard language of Arabic Newswire. It used to reduce the noise in the texts, through remove irrelevant or not important words such as stop words, prepositions, punctuation marks, digits from Arabic texts. After that, replace some Arabic characters into other characters to be more understandable and readable by computer.

Text normalization is applying on several natural language texts. It represents a task to transfer the inconsistence text to be more consistency. In the Arabic language was used normalization to remove the diacritics marks, and normalize the other specific characters. **Tokenization** is a process to divide the plain text into tokens to remove the noise from the text. After that sent it into morphological analyzer to continue the processing [3]. **Stop-words removal** process is to remove the frequent Arabic words that insignificant words or aren't carry important meaning.

6. Discussion

The proposed approach has been applied on a sample of 40- documents, and 3- queries. The Figure is shown below contains the related top documents that retrieved using both proposed method with light-stemmer, and without stemming.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Recall	Precision with-Light Stemmer	Precision without-Light Stemmer
0	1	1
0.1	0.962237954	0.892134
0.2	0.927821344	0.810019
0.3	0.926545433	0.7817921
0.4	0.913451561	0.7224335
0.5	0.901349563	0.6691001
0.6	0.883913395	0.471761696
0.7	0.868433115	0.398249601
0.8	0.855947489	0.305737507
0.9	0.830516666	0.216225412
1	0.801558441	0.156713317

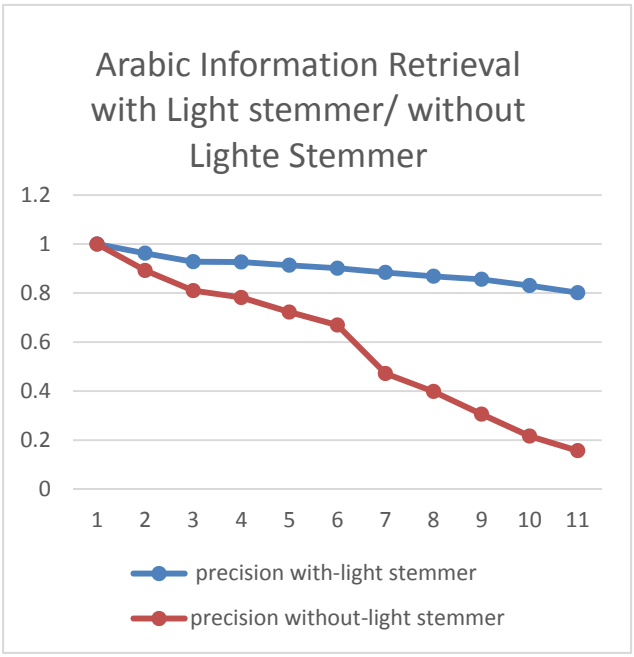

Figure 1: The performance of Arabic retrieval using Light stemmer, and without stemmer

The results of this proposed method using light-stemmer show a good measurement better that without stemming. The interpolated precision to explain of improvement of the performance Arabic information retrieval using the proposed method.

7. Conclusion

This study based keywords- searching through light stemmer to retrieve the information needs from spread Arabic contents in the web. It shows the effectiveness of light stemming on Arabic information retrieval. Light stemming helps the retrieval systems, and search engines through enhancing the search performance of these systems. This study confirms that the light stemming helps to match many of the words that share in the origin of the word or root by deleting the prefix and suffixes, thus allowing the matching of the most different words in the text.

References

1. Abdelali, A., Cowie, J., Farwell, D., Ogden, B., & Helmreich, S. (2003). *Cross-language information retrieval using ontology*. Paper presented at the Proc. of the Conference TALN 2003.
2. Al-Khalifa, H., & Al-Wabil, A. (2007). *The Arabic language and the semantic web: Challenges and opportunities.* Paper presented at the The 1st int. symposium on computer and Arabic language.
3. Al-Taani, A. T., & Al-Rub, S. A. (2009). A rule-based approach for tagging non-vocalized Arabic words. *Int. Arab J. Inf. Technol., 6*(3), 320-328.
4. Al-Taani, A. T., Msallam, M. M., & Wedian, S. A. (2012). A top-down chart parser for analyzing arabic sentences. *Int. Arab J. Inf. Technol., 9*(2), 109-116.
5. Elabd, E., Alshari, E., & Abdulkader, H. (2015). Semantic Boolean Arabic Information Retrieval. *arXiv preprint arXiv:1512.03167*.
6. Haav, H.-M., & Lubi, T.-L. (2001). *A survey of concept-based information retrieval tools on the web*. Paper presented at the Proceedings of the 5th East-European Conference ADBIS.
7. Hammo, B., Abu-Salem, H., & Lytinen, S. (2002). *QARAB: A question answering*

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

system to support the Arabic language. Paper presented at the Proceedings of the ACL-02 workshop on Computational approaches to semitic languages.

8. Larkey, L., Ballesteros, L., & Connell, M. (2007). Light stemming for Arabic information retrieval. *Arabic computational morphology*, 221-243.
9. Noordin, M. F., & Othman, R. (2006). *An information retrieval system for Quranic texts: a proposed system design*. Paper presented at the Information and Communication Technologies, 2006. ICTTA'06. 2nd.
10. Samy, D., Moreno-Sandoval, A., Bueno-Díaz, C., Salazar, M. G., & Guirao, J. M. (2012). *Medical Term Extraction in an Arabic Medical Corpus*. Paper presented at the LREC.
11. Vallet, D., Fernández, M., & Castells, P. (2005). *An ontology-based information retrieval model*. Paper presented at the European Semantic Web Conference.
12. Wan, R., Anh, V. N., & Mamitsuka, H. (2007). *Passage Retrieval with Vector Space and Query-Level Aspect Models*. Paper presented at the TREC.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500